# Customs Tariff Classification of Organic Chemicals With Data Mining Techniques

Siriwut Kongnakorn
*College of Innovation*
*Thammasat University*
Bangkok, Thailand
siriwut.kongnakorn@gmail.com

Tinnarat Aromsuk
*Collage of Innovation*
*Thamasat University*
Bangkok , Thailand
tinnarat.aro@dome.tu.ac.th

Wasinee Noonpakdee
*Collage of Innovation*
*Thamasat University*
Bangkok , Thailand
wasinee@citu.tu.ac.th

***Abstract*— Harmonized System (HS Code) is a widely used classification system for commodities and facilitating international trade with 97 chapters according to commodity types. Classifying some of the 97 chapters of the HS Code can be a time-consuming and demanding process. In particular, the classification of organic chemicals belonging to HS Code Chapter 29 is especially challenging and requires specialized knowledge. This research aims to examine the classification of the first four digits of HS Code Chapter 29 of organic chemicals (headings 29.01-29.35) using IUPAC names, a globally recognized chemical nomenclature and terminology. In this study, two machine learning models, K-Nearest Neighbor (K-NN) and Support Vector Machines (SVMs), were employed and their parameters were optimized for maximum accuracy. Results showed that the SVM model with optimized parameters achieved the highest accuracy. The results of this study can be applied to improve the efficiency and accuracy of customs operations in the future.**

***Keywords*— *Harmonized System, Classification Techniques, IUPAC Name, Text Mining, Customs Tariff Classification***

## I. Introduction

Due to the wide variety of products of each country, the World Customs Organization (WCO) has developed a product classification system for facilitating international trade called the Harmonized System (HS Code). HS Code contains 6-digit numeric codes and their goods descriptions which are split into 21 Sections or 97 Chapters according to product groups. [1, 2]

Goods in Chapter 29 of the HS Code are organic chemicals. The classification of organic chemicals that belong to Chapter 29 of the HS Code presents a significant challenge for customs officers, as it requires specialized knowledge of organic chemistry and can be time-consuming. Previous studies have primarily focused on the classification of all 97 chapters of the HS Code and relied heavily on data from goods descriptions [3-5]. As a result, there is a need for more advanced and targeted research to improve the accuracy and efficiency of HS Code classification for organic chemicals in Chapter 29.

For Chapter 29 of the HS Code, the heading of 29.01-29.35 is based on the chemical structures of organic compounds. Organic compounds have a worldwide chemical nomenclature and terminology called the IUPAC system (International Union of Pure and Applied Chemistry). Several studies have applied data mining techniques to the IUPAC names of chemicals, such as a study that used a transformer-based model to classify chemical structures from chemical compound names (both IUPAC name and non-IUPAC name) [6] and another study that developed a transformer-based artificial neural network to convert between chemical structures and IUPAC names [7]. Other studies have explored text analysis using various machine learning algorithms, such as Turkish news classification using Naïve Bayes and SVM [8], Vietnamese news classification using Neural Network and SVM [9], and short text sentiment classification using Bi-LSTM [10].

This research aims to investigate the HS Code classification process for organic chemicals in Chapter 29, focusing on heading 29.01-29.35. The study will utilize K-Nearest Neighbor (K-NN) and Support Vector Machines (SVMs) as classification models to predict the HS Code from the IUPAC names of organic chemicals. Furthermore, this study will compare the performance of these models to identify the most effective approach. The results can be applied to assist customs operations with efficiency and accuracy in the future.

## II. Literature Review

### A. HS Code

Harmonized System (HS Code) is developed by the World Customs Organization (WCO) as a product classification system for facilitating international trade. HS Code contains 6-digit numeric codes and their goods descriptions which are split into 21 Sections or 97 Chapters according to product groups. Each Chapter consists of one or more headings and each heading consists of one or more subheadings. Chapter, heading and subheading are identified by the first two digits, first four digits, and six digits of the HS Code, respectively [1, 2]. An example of HS Code Chapter 29 (heading 29.01) [2] is shown in figure 1.

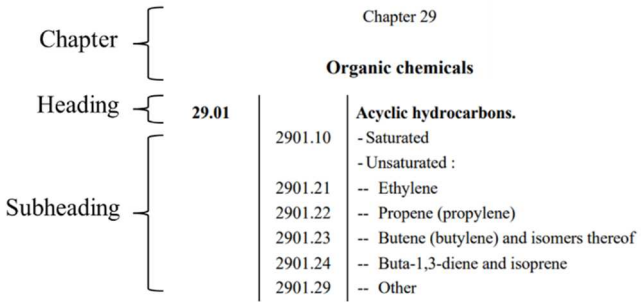


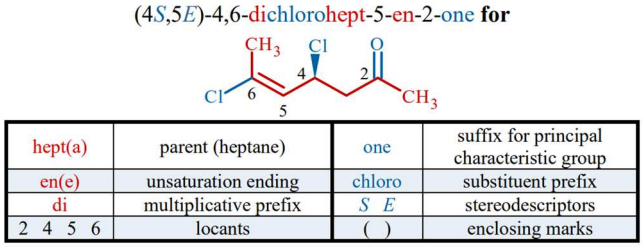

Fig. 1. Example of HS Code Chapter 29 (heading 29.01) [2]

Goods in Chapter 29 of the HS Code are organic chemicals. Chapter 29 can be divided into three groups. The first one, based on the chemical structures of organic compounds, contains heading 29.01-29.35. The second one, belonging to certain groups having physiological or biological functions, contains heading 29.36-29.41. The last one is

residual heading 29.42, other organic compounds other than those two groups [2].

### B. IUPAC system

Organic compounds have a worldwide chemical nomenclature and terminology called the IUPAC system (International Union of Pure and Applied Chemistry). IUPAC nomenclature of organic chemicals has a set of guidelines to create systematic names (IUPAC name) for chemical compounds from their chemical structures. Therefore, the IUPAC names of organic compounds are related to their chemical structures [11] as shown in figure 2.




| hept(a) | parent (heptane) | one | suffix for principal characteristic group |
|---|---|---|---|
| en(e) | unsaturation ending | chloro | substituent prefix |
| di | multiplicative prefix | S E | stereodescriptors |
| 2 4 5 6 | locants | ( ) | enclosing marks |

Fig. 2. Example of IUPAC name and chemical structure of an organic compound [11]

### C. Data Mining techniques

In past research, data mining techniques have been applied to the IUPAC names of chemicals. For example, Krasnov et al. [7] create a transformer-based model to convert chemical structure data back to the IUPAC name by using an artificial neural network to compare performance with the rule-based solution method. The study's dataset comes from the PubChem database, which collects large amounts of chemical data. In the preparation of IUPAC data, Krasnov et al. [7] developed their rule-based tokenization method to tokenize IUPAC data whose data type is text. It sorts out the components of an IUPAC name: prefixes (e.g., "-oxy", "-hydroxy", "-di", "-tri", etc.), suffixes (e.g., "-one", "-ol", etc.), numbers and special symbols including stereochemical symbols (e.g. “(”, “)”, “[”, “]”, “-”, “N”, “R or S”, “E or Z”, etc.).

Another research by Omote et al. [6] developed a model for transforming chemical names in IUPAC and non-IUPAC systems into chemical structures by using a sequence-to-sequence neural network. In addition, atomic number loss function computation techniques are used together with multi-task learning to optimize the model. Then the obtained model was compared with the rule-based method. In this research, the dataset from the PubChem database was used. The IUPAC name data was tokenized using OPSIN (Open Parser for Systematic IUPAC Nomenclature) [12] which is a free software used to convert chemical names in the IUPAC system to text-based data representing chemical structures such as SMILES and InChI.

In text analysis or text mining, Gürcan [8] classified Turkish news into 5 classes (sport, health, technology, politics, and Economy) using supervised learning algorithms (e.g., Naïve Bayes, K-Nearest Neighbor (K-NN), Support Vector Machine (SVM), and Decision Trees (J48), etc.) and compared the performance of each model. Another similar study by Phuoc Vinh and Kha [9] classified Vietnamese news into 30 classes using neural networks for feature extraction to reduce the dimensionality of the data and using SVMs as a classification technique.

In addition, the research of Yang et al. [10] classified short sentimental texts using feature expansion techniques, followed by Bi-directional Long Short-Term Memory (Bi-LSTM), a type of deep neural network algorithm, to extract the features of the data. Then pass it through a SoftMax Classifier to predict the result of text classification.

Several studies have been conducted on HS Code classification. The study in [3] show HS Code classification from goods descriptions of Indonesian import data. For the classification of the first four digits of HS Code, Random Forest leads the other classifier with accuracy, precision, recall, and F1-score of 70.73%, 85.50%, 76.95%, and 79.60%, respectively.

In [4], a neural machine translation model is used with the integration of hierarchical loss (NMT-HL) to classify 6-digit HS codes by using a dataset from DHL, a logistics services company. The dataset contains item description, origin, destination, origin airport, and destination airport. The model shows an accuracy of 85% but a recall of around 29%. As the model is trained to increase in accuracy, the model's recall will decrease.

The HS Code is predicted in [5] using the customers’ input goods descriptions. For the classification of the first four digits of HS Code, The linear support vector machine model has the highest accuracy of 84.58%. Moreover, the model has precision, recall, and F1-score of 96.35%, 70.55%, and 81.46%, respectively.

From the literature review, therefore, OPSIN software is used for tokenization with IUPAC name data as in the work of Omote et al. [6] because OPSIN is free software and can perform tokenization with chemical name data in the IUPAC system. It saves both time and resources to rebuild the new tokenization model. The algorithm used to classify the HS Code of an organic compound from the IUPAC name will select data classification techniques to compare performance to find the best model. In this research, K-Nearest Neighbors (K-NN) and Support Vector Machines (SVMs) techniques will be used to predict the HS Code from the IUPAC names of organic compounds.

## III. Research Method

In this study, there are four main procedures: data collection, data preparation, data mining, and model evaluation. An overview of the research procedure is shown in figure 3.

### A. Data Collection

This research used three datasets comprising 9660 records with attributes such as IUPAC name, non-IUPAC name, and the latest edition of the HS Code for organic compounds (2022). The class label was divided into 35 classes based on headings 29.01-29.35 in Chapter 29 of the HS Code. There are the three datasets as follows:

*1) Dataset 1:* HS Code classification of pharmaceutical chemicals dataset called International Nonproprietary Name (INN) or INN Table from the WCO website, containing 5467 records.

*2) Dataset 2:* Customs tariff data with statistics code 2022 edition (only heading 29.01-29.35) from the Integrated Tariff Database of Tariff Structure Section, Customs Tariff

Division, Thai Customs Department, containing 1801 records.

*3) Dataset 3:* IUPAC name dataset (only HS Code heading 29.01-29.35) from the PubChem website, containing 2392 records.

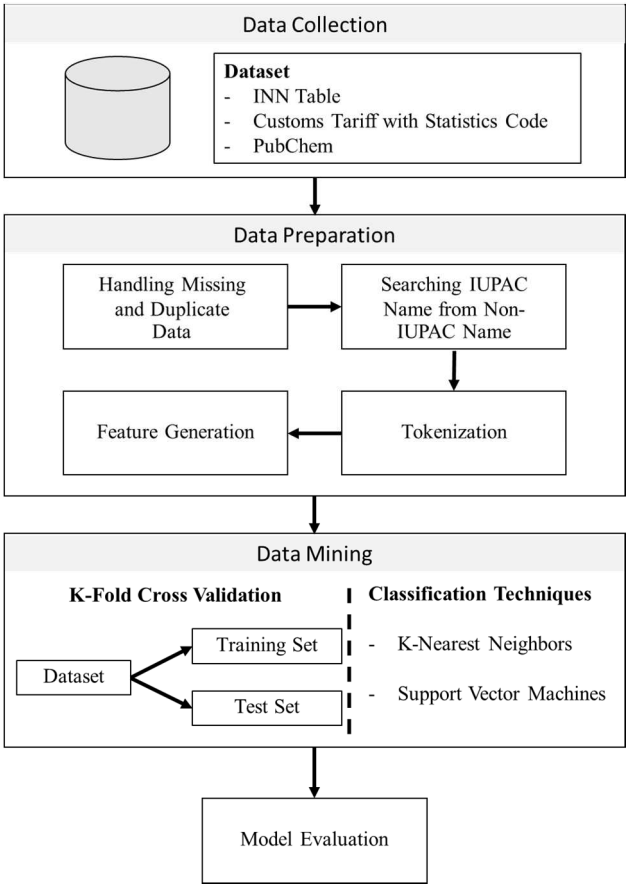



Fig. 3. Research method

## B. Data Preparation

*1) Handling Missing and Duplicate Data:* All columns from all datasets are removed except those related to chemical names (IUPAC names and non-IUPAC names), and HS Code. After that, select only records whose HS codes are in heading 29.01-29.35 and remove records other than those. Then, detect and remove records that contain missing values or duplicate data from all datasets.

| | HS Code | Name |
|---|---|---|
| 1205 | 29.31 | 2-chlorovinyldichloroarsine |
| 1293 | 29.31 | 2-chlorovinyldichloroarsine |
| 1294 | 29.31 | 2-chlorovinyldichloroarsine |
| 1204 | 29.31 | 2-chlorovinyldichloroarsine |
| 1250 | 29.31 | Benzenearsonic acid |
| ... | ... | ... |
| 1216 | 29.31 | Trichlorotrivinylarsine |
| 1304 | 29.31 | Vinylchloroarsine |
| 1218 | 29.31 | Vinylchloroarsine |
| 1217 | 29.31 | Vinylchloroarsine |
| 1303 | 29.31 | Vinylchloroarsine |

[195 rows x 2 columns]

Fig. 4. Example duplicate data in Dataset 2

*2) Searching IUPAC Name from Non-IUPAC Name:* Dataset 1 and 2 have the attribute "Product name" which contains chemical names but is not IUPAC names. In this process, the values in the attribute of those datasets are used to fetch IUPAC names from the PubChem database by coding Python script from Library named PubChemPy [13]. In the case of Dataset 3, this step is not required because Dataset 3 is collected from the PubChem database and already has the IUPAC name attribute. After that, all the datasets are merged and missing values or duplicate data are removed before the tokenization process.

*3) Tokenization:* In general, tokenization of English texts or sentences is often used to split text from the spacing between words. However, this research uses the IUPAC name, which is a text with a specific structure. Preparation of the IUPAC name requires splitting and collecting the prefixes and suffixes of the IUPAC name since they represent the chemical structure of the chemical compound in figure 5. In this paper, IUPAC names were tokenized with a Java script from OPSIN (Open Parser for Systematic IUPAC Nomenclature) [12].

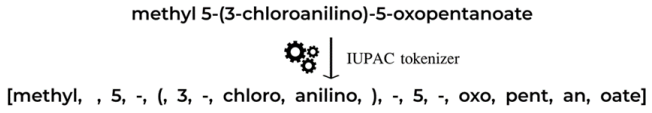



Fig. 5. Example of the result of Tokenization with an IUPAC name [7].

*4) Feature Generation:* Creating a bag of words from all the words in the tokenized IUPAC name data, entering a number for each word found in each record of the dataset. There are several ways to add numbers as follows.

*a) Binary Term Occurrence:* Replaces 0 when the word in the bag of words is not found in the record of the dataset and 1 when found.

*b) Term Occurrence:* Count the number of individual words found in the record of the dataset.

*c) Term Frequency (TF):* Count the number of individual words found in the record of the dataset and divide by the number of all words found in that record.

*d) Term Frequency-Inverse Document Frequency (TF-IDF):* TF-IDF can be calculated in the equation below.

$$IDF(W) = log(N/n) \quad (1)$$

$$TF\text{-}IDF = TF*IDF(W) \quad (2)$$

Where $W$ is the word, $N$ is the number of all records in the dataset and $n$ is the number of records containing the word $W$.

## C. Data Mining and Model Evaluation

In this research, RapidMiner Studio was used for text mining by splitting the dataset into the training set and test set using k-fold cross-validation with k = 5 and performing text analysis by using classification techniques with parameter adjustments to find the model with the best performance as follows:

*1) Support Vector Machines (SVMs):* A technique to create a plane to divide data into two groups by trying to create a plane with the most margin [9].

*2) K-Nearest Neighbors (K-NN):* A technique that compares the predicted data from the test set with the training set. If the predicted data is similar to data in the Training Set, the predicted data will be classified into the same class as that of the similar data.

In K-NN, k means the number of data in the training set that is compared to the data to be predicted. For example, if k is equal to 5, the K-NN algorithm compares the data to be predicted with the data in the training set to find 5 records in the training set that are most similar to the predicted data.

There are several ways to calculate data similarity or distance (e.g., Euclidean, Manhattan, Hamming, etc.) which can be calculated using the equation below, where *x* and *y* are the data values in the training set and the data values to be predicted [14].

$$Euclidean\ Distance = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \quad (3)$$

$$Manhattan\ Distance = \sum_{i=1}^{k}|x_i - y_i| \quad (4)$$

In the research, Accuracy, Precision, Recall, and F1-Score values are used to evaluate model performance. The equations used to calculate are as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

$$F1 - Score = 2 \cdot \frac{Precision \cdot Recall}{Precision \times Recall} \quad (8)$$

Where TP (True Positive): The number of values that the model correctly predicts as positive.

TN (True Negative): The number of values that the model correctly predicts as negative.

FP (False Positive): The number of values that the model incorrectly predicts as positive.

FN (False Negative): The number of values that the model incorrectly predicts as negative.

## IV. Result and Discussion

This study handles missing and duplicate data in the datasets, and adds IUPAC names by searching for them from non-IUPAC names. Afterwards, the datasets are merged into one. Because the dataset has the problem of imbalanced data, they are reduced by a random method for some classes with many records. Finally, there are 4,861 rows of data left, divided into 35 classes as shown in figure 6.

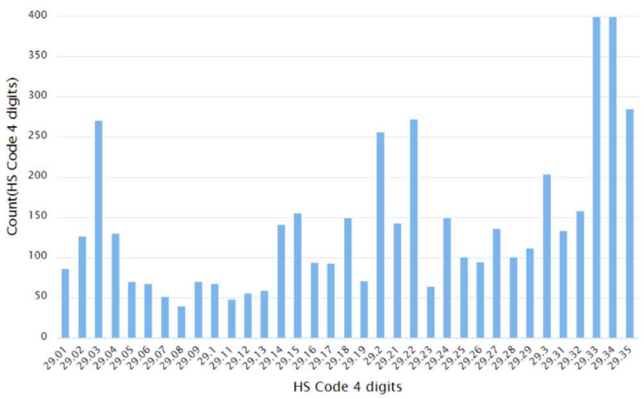



Fig. 6. Number of records in each class from class 29.01-29.35 of the dataset after reduced by a random method.

Next, the IUPAC Name attribute data was tokenized using a Java script from OPSIN. The tokens of each record were collected and joined using the "|" symbol. The 10 example records are illustrated in Table 1. However, a few records (2-3 records) could not be tokenized due to their long and complex IUPAC names, so they were removed before proceeding to the next step.

After that, the dataset was imported into RapidMiner Studio for feature generation and divided into the training set and test set using the k-fold cross-validation technique with k = 5. Then, the dataset was analyzed using K-Nearest Neighbor (K-NN) and Support Vector Machines (SVMs) by changing the text vectorization method (e.g., Binary Term Occurrence, Term Occurrence, Term Frequency (TF) and Term Frequency-Inverse Document Frequency (TF-IDF)). For the parameter settings of K-NN, using k = 5, the Euclidean distance method is used to determine the distance between the data. For the parameter settings of SVMs, the Support Vector Machine (LibSVM) operator with default parameter settings in RapidMiner Studio is used, whose kernel type is rbf. The results are shown in Table 2 and 3.

TABLE I. 10 Example Records of the Dataset after Tokenization

| IUPAC Name | Tokens | HS Code 4 digits |
|---|---|---|
| buta-1,3-diene | but\|a\|-\|1,3-\|di\|ene | 29.01 |
| 4-nitrobenzenesulfonic acid | 4-\|nitro\|benzen\|e\|sulfon\|ic acid | 29.04 |
| butane-1,4-diol | but\|ane\|-\|1,4-\|di\|ol | 29.05 |
| 1,1-diethoxypentane | 1,1-\|di\|eth\|oxy\|pent\|ane | 29.11 |
| 4-phenylbutan-2-one | 4-\|phenyl\|but\|an\|-\|2-\|one | 29.14 |
| 2,2-dimethylpropanedioic acid | 2,2-\|di\|meth\|yl\|prop\|ane\|di\|o\| ic acid | 29.17 |
| 2-chloro-N-hex-5-ynylacetamide | 2-\|chloro\|-\|N-\|hex\|-\|5-\|yn\|yl\|acet\|amide | 29.21 |
| 3-bromobenzonitrile | 3-\|bromo\|benz\|o\|nitrile | 29.26 |
| bis(2-chloroethylamino) phosphinic acid | bis\|(\|2-\|chloro\|eth\|yl\|amino\|)\| phosphin\|ic acid | 29.29 |
| 6-phenylpteridine-2,4,7-triamine | 6-\|phenyl\|pteridin\|e\|-\|2,4,7-\|tri\|amine | 29.33 |

TABLE II. CLASSIFICATION RESULTS FROM K-NN WITH VARIOUS TEXT VECTORIZATION METHODS

| Text Vectorization method | Accuracy (%) | Precision (%) | Recall (%) | F1-score (%) |
|---|---|---|---|---|
| Binary Term Occurrence | 55.75 | 62.56 | 59.12 | 60.79 |
| Term Occurrence | 56.97 | 63.36 | 59.17 | 61.19 |
| TF | 60.20 | 67.30 | 61.16 | 64.08 |
| TF-IDF | 62.75 | 65.45 | 63.96 | 64.70 |

TABLE III. CLASSIFICATION RESULTS FROM SVMS WITH VARIOUS TEXT VECTORIZATION METHODS

| Text Vectorization method | Accuracy (%) | Precision (%) | Recall (%) | F1-score (%) |
|---|---|---|---|---|
| Binary Term Occurrence | 9.03 | 0.56 | 3.14 | 0.95 |
| Term Occurrence | 14.08 | 3.67 | 5.89 | 4.52 |
| TF | 8.44 | 0.28 | 2.94 | 0.51 |
| TF-IDF | 8.83 | 0.41 | 3.07 | 0.72 |

According to Table 2 and 3, K-NN with TF-IDF has the highest accuracy at 62.75%. When replacing text vectorization techniques with TF, Term Occurrence, and Binary Term Occurrence, accuracy decreases slightly respectively, whereas SVMs haven't high accuracy which is less than 15%. In the case of SVMs with the RBF kernel, the parameters C and gamma were set to their default values (C=0, gamma=0). When the gamma value is too low, it causes the shape of planes that separate the data in SVMs to have low complexity, and not fit the shape of the data. Therefore, it causes the model to have low performance.

In addition, the parameters were adjusted using TF-IDF as a text vectorization technique. For K-NN, the k-value was adjusted to 3, 5, 7, and 9, and the parameters including the k-value and the distance measurement method were optimized using Optimize Parameters (Grid) operator as shown in Table 4. For SVMs, modify kernel type value (e.g., linear, rbf, poly, sigmoid) and optimize parameters (e.g., kernel type, gamma, and C value) using Optimize Parameters (Grid) operator which shows performance in Table 5.

TABLE IV. CLASSIFICATION RESULTS FROM K-NN WITH TF-IDF AND OTHER VARIOUS PARAMETER SETTINGS

| Parameter Settings | Accuracy (%) | Precision (%) | Recall (%) | F1-score (%) |
|---|---|---|---|---|
| k = 3 | 63.02 | 66.04 | 64.7 | 65.36 |
| k = 5 | 62.75 | 65.45 | 63.96 | 64.70 |
| k = 7 | 63.22 | 66.00 | 63.88 | 64.92 |
| k = 9 | 63.22 | 66.58 | 63.43 | 64.97 |
| Optimize parameter (best accuracy at k=4, Manhattan distance) | 63.70 | 67.45 | 66.41 | 66.93 |

TABLE V. CLASSIFICATION RESULTS FROM SVMS WITH TF-IDF AND OTHER VARIOUS PARAMETER SETTINGS

| Parameter Settings | Accuracy (%) | Precision (%) | Recall (%) | F1-score (%) |
|---|---|---|---|---|
| kernel type=linear | 77.13 | 79.06 | 75.58 | 77.28 |
| kernel type=poly, degree=3 | 68.86 | 77.83 | 65.95 | 71.40 |
| kernel type=rbf | 8.83 | 0.41 | 3.07 | 0.72 |
| kernel type=sigmoid | 77.13 | 79.06 | 75.58 | 77.28 |
| Optimize parameter (best accuracy at kernel type=linear, gamma=1.52, C=2.98) | 79.07 | 81.61 | 78.78 | 80.17 |

According to Table 4 and 5, K-NN with TF-IDF and four modified k values (k=3, 5, 7, 9) have similar accuracy of about 63%. In addition, K-NN with Optimize Parameters (Grid) operator, at k = 4 and using Manhattan distance as a distance function, has an accuracy of 63.70%, which is the highest value when compared to the same model. In the case of SVMs with TF-IDF, the kernel type being linear or sigmoid has high accuracy at 77.13%, followed by the kernel type being polynomial at 68.86% and the minimum accuracy for kernel type being the radial basis function (rbf) at 8.83%. Moreover, SVMs with Optimize Parameters (Grid) operator has an accuracy of 79.07% for kernel type=linear, gamma=1.52, and C=2.98, which is the highest accuracy among all models.

Comparing the performance of this study to previous research on HS code classification, the performance of this study is similar to the study in [3], which used goods descriptions of Indonesian import data to classify HS Codes with an accuracy of 70.73%, precision of 85.50%, recall of 76.95%, and F1-score of 79.60%. In contrast, the study by [4] using a neural machine translation model with the integration of hierarchical loss (NMT-HL) to classify 6-digit HS codes by using a dataset from DHL, achieved an accuracy of 85% but with low recall. The study by [5] had the highest accuracy of 84.58% with precision of 96.35%, recall of 70.55%, and F1-score of 81.46%. However, these previous studies focused on classifying all 97 chapters of the HS Code using goods descriptions, while this research focuses on Chapter 29 and utilizes IUPAC names of organic chemical products for classification.

## V. CONCLUSION

This paper presents the classification of the first four digits of HS Code Chapter 29, focusing on headings 29.01 to 29.35, using the IUPAC names of organic chemicals. The classification was performed using K-Nearest Neighbor (K-NN) and Support Vector Machines (SVMs). The models were adjusted for various parameters and found that SVMs with Optimize Parameters (Grid) operator (at kernel type=linear, gamma=1.52, and C=2.98) has the highest accuracy of 79.07%.

The results of this study could be utilized to improve the efficiency of customs work and classification procedures in the future. Further research could be conducted by exploring other imbalanced data solutions and applying other machine learning models such as Artificial Neural Network (ANN), Decision Tree, or Naive Bayes to enhance the performance of the classification models.

## REFERENCES


[1] World Customs Organization. (2022). *What is the Harmonized System (HS)?* Available: http://www.wcoomd.org/en/topics/nomenclature/overview/what-is-the-harmonized-system.aspx

[2] World Customs Organization. (2022). *HS Nomenclature 2022 edition*. Available: http://www.wcoomd.org/en/topics/nomenclature/instrument-and-tools/hs-nomenclature-2022-edition/hs-nomenclature-2022-edition.aspx

[3] I. G. Y. Paramartha, I. Ardiyanto, and R. Hidayat, "Developing Machine Learning Framework to Classify Harmonized System Code. Case Study: Indonesian Customs," presented at the 2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT), 2021.

[4] X. Chen, S. Bromuri, and M. van Eekelen, "Neural Machine Translation for Harmonized System Codes Prediction," *Icmlt 2021,* pp. 158–163, 2021.

[5] F. Altaheri and K. Shaalan, "Exploring Machine Learning Models to Predict Harmonized System Code," pp. 291-303, 2020.

[6] Y. Omote, K. Matsushita, T. Iwakura, A. Tamura, and T. Ninomiya, "Transformer-based Approach for Predicting Chemical Compound Structures," Suzhou, China, 2020, pp. 154-162: Association for Computational Linguistics.

[7] L. Krasnov, I. Khokhlov, M. V. Fedorov, and S. Sosnin, "Transformer-based artificial neural networks for the conversion between chemical notations," *Sci Rep,* vol. 11, no. 1, p. 14798, Jul 20 2021.

[8] F. Gürcan, "Multi-Class Classification of Turkish Texts with Machine Learning Algorithms," presented at the 2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), 2018.

[9] T. N. Phuoc Vinh and H. H. Kha, "Feature Extraction Using Neural Networks for Vietnamese Text Classification," presented at the 2021 International Symposium on Electrical and Electronics Engineering (ISEE), 2021.

[10] C. Yang, W. Zheng, Y. Xiao, and C. Dong, "A short text sentiment classification method based on feature expansion and bidirectional neural network," presented at the 2021 International Conference on Big Data Analysis and Computer Science (BDACS), 2021.

[11] K.-H. Hellwich, R. M. Hartshorn, A. Yerin, T. Damhus, and A. T. Hutton, "Brief guide to the nomenclature of organic chemistry (IUPAC Technical Report)," *Pure and Applied Chemistry,* vol. 92, no. 3, pp. 527-539, 2020.

[12] D. Lowe. (2022). *OPSIN: Open Parser for Systematic IUPAC nomenclature*. Available: https://opsin.ch.cam.ac.uk/

[13] M. Swain. (2014). *PubChemPy documentation*. Available: https://pubchempy.readthedocs.io/en/latest/index.html

[14] S. Jain, D. S. C. Jain, and D. S. K. Vishwakarma, "Analysis of Text Classification with various Term Weighting Schemes in Vector Space Model," *International Journal of Innovative Technology and Exploring Engineering,* vol. 9, no. 10, pp. 390-393, 2020.